IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

AI-Powered Video and Image Procssing for Video Meetings

Abhishek Mohite, Ayush Rathod, Gaurang Arora, Nishant Khanderao, Prof. D. H. Kulkarni
Smt. Kashibai Navale College of Engineering ,Vadgaon, Pune Savitribai Phule Pune University

Abstract: *This project presents an AI-powered system for real- time video and image processing in virtual meetings, aimed at enhancing security, engagement, and integrity. It integrates deep learning techniques including YOLO for object detection, face recognition, facial landmark analysis, and MediaPipe-based head pose estimation. The system detects behaviors such as gaze diversion, mouth movement, multiple person presence, and spoofed identities using color histogram-based spoof detection. Alerts are generated for suspicious activities in real time, making it ideal for online proctoring, virtual interviews, and remote monitoring. The architecture ensures low latency and high accuracy, supporting intelligent and trustworthy remote interactions.*

Keywords: Online proctoring system, Education, Authentication, Abnormal behavior detection

I. INTRODUCTION

The transition to remote communication has significantly increased the dependence on virtual meeting platforms across education, corporate, and professional domains. While these platforms enable seamless connectivity, they fall short in ensuring participant authenticity, attention, and behavioural accountability—particularly in high-stakes settings such as online examinations, interviews, and corporate reviews. To address these challenges, this research introduces an AI-powered video and image processing framework designed for real-time behavioural monitoring and security in virtual meetings.

The proposed system integrates multiple advanced computer vision techniques to analyse live webcam feeds and detect anomalies. Object detection is performed using the YOLOv8 model to identify the presence of unauthorized persons or banned items such as mobile phones, books, or multiple users. Facial recognition ensures participant authentication, while Dlib-based landmark detection tracks eye movement, blinking, and mouth activity to infer attentiveness or verbal activity. MediaPipe-based head pose estimation is employed to identify deviations in focus, and spoof detection is achieved through a machine learning classifier trained on YCrCb and LUV color histograms.

The entire system operates in real time, delivering low-latency alerts when abnormal behaviors are consistently detected. This makes the solution particularly suitable for online proctoring, remote interviews, and secure conferencing environments. By combining object-level and behavior-level analysis, the proposed framework advances the state of AI-driven virtual interaction, providing an intelligent and automated layer of oversight that enhances the reliability and integrity of remote communication.

II. LITERATURE REVIEW

Existing Proctoring Systems

S.Prathish et al. [1] used the model-based head pose estima- tion method and the audio-based detection method to com- plete the test abnormal behavior detection. However, the ac- curacy rate of the head pose estimation of this method is not high enough, and the use of a microphone to collect sound can infringe the relevant privacy of examinees. Moreover, the abnormal behavior detection process does not consider eye tracking and mouth movement analysis.

In [2], the authors proposed a multimedia analytics sys- tem for online exam proctoring. With the captured videos and audio, they extract low-level features from six basic compo- nents: text detection, user verification, active window detec- tion, speech detection, gaze estimation, and phone detection. These features are then processed in a temporal window to

DOI: 10.48175/IJARSCT-27905

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

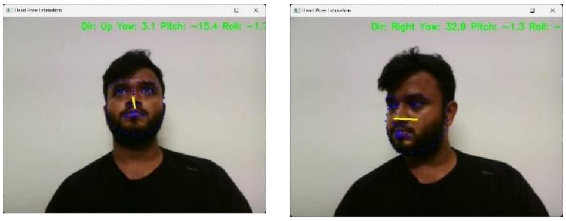
acquire high-level features and then used for cheat detection. However, the system is not feasible as it requires the exami- nee to have a wearcam. Moreover, the solution does not have a face spoofing feature for user authentication.

Hu et al. [3] proposed a system that uses an image-based head pose estimation model and mouth movement analysis to discriminate the abnormal behavior of the examinee the online examination. However, the system does not take eye-tracking functionality into consideration for analyzing the abnormal behavior of the examinee

Head Pose Estimation

Head pose estimation plays a pivotal role in assessing user attention and behaviour during virtual interactions, especially in proctored or monitored environments. It involves determining the orientation of a user's head in three-dimensional space relative to the camera, which provides critical insights into whether the participant is focused on the screen or distracted. In the proposed system, head pose estimation is implemented using a combination of MediaPipe's facial mesh detection and the Perspective-n- Point (PnP) algorithm via OpenCV. MediaPipe extracts 3D facial landmarks with high precision, from which a subset of key reference points—such as the nose tip, eye corners, and mouth corners—is selected. These are mapped to a standardized 3D face model to solve for rotation and translation vectors using the solvePnP algorithm.

The resulting vectors are converted to Euler angles (yaw, pitch, and roll), which indicate the directional orientation of the user's head. Based on these angles, the system classifies head movement into categories such as "Left," "Right," "Up," "Down," or "Center." Persistent deviations from the "Center" pose can signify disengagement, possible malpractice, or interaction with unauthorized materials. This module, when combined with gaze estimation and facial activity tracking, provides a robust behavioural analysis toolkit. The real-time, non-intrusive nature of this approach ensures continuous monitoring without affecting user experience. Head pose estimation thereby enhances the system's ability to ensure integrity and attention in remote meetings or assessments, making it a core component of AI- driven proctoring frameworks.



DATASET

For the development and validation of the face spoof detection module, we utilized the CASIA Face Anti-Spoofing Database (CASIA-FASD) [9], a widely adopted benchmark for evaluating the performance of anti-spoofing algorithms. The dataset consists of 600 video sequences collected from 50 subjects under three different imaging qualities (low, normal, and high resolution). Each subject is recorded under genuine (live) and attack conditions, including printed photos, replayed videos, and warped images. The dataset simulates realistic spoofing scenarios to train robust models capable of generalizing across diverse attack types. Each video sequence is labeled with ground truth for binary classification (live or spoof), and face bounding boxes are provided for consistency during preprocessing.

To enhance the diversity of the dataset and mitigate the risk of model overfitting, a range of data augmentation techniques was applied. These include geometric transformations such as horizontal flipping and random scaling, as well as color-based perturbations like brightness and contrast jittering. Additionally, noise-based augmentations—such as Gaussian noise, salt-and- pepper noise, and random occlusion—were introduced to mimic environmental variations. These augmentation strategies ensure the model remains invariant to minor appearance changes and lighting inconsistencies. Histogram-based features from YCrCb and LUV color spaces were extracted to train a machine learning classifier for spoof detection. This pipeline enables the system to reliably distinguish between live and spoofed faces in real-time video streams, ensuring secure and tamper-resistant virtual authentication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

III. PROPOSED SYSTEM

System Architecture

This work proposes a system that uses a webcam to monitor examinees during the online examination. The system archi- tecture is shown in Fig. 3. After the camera captures the frame, banned item detection and person detection are per- formed. If one person is detected in the frame, then face de- tection is performed. The detected face is input to face spoof- ing, face verification, face landmark detection, and head pose estimation models. The detected landmarks from the land- mark detection model are input to mouth movement analysis and eye-tracking. We analyze the output from all models to conclude whether the examinee is cheating or not. Table 2 contains all the functionalities present in our proposed system.

Table 2. Functionalities in our system

Category	Functionality	Technique
Base	Person Detection and Counting Object Detection Face Detection	Pre-trained Pre-trained Pre-trained
Authentication	Face Verification Face Spoofing	Pre-trained Pre-trained
Abnormal Behavior Detection	Image-based Head Pose Estimation Eye Tracking Mouth Movement Analysis	Trained from Scratch Image-Processing Image-Processing

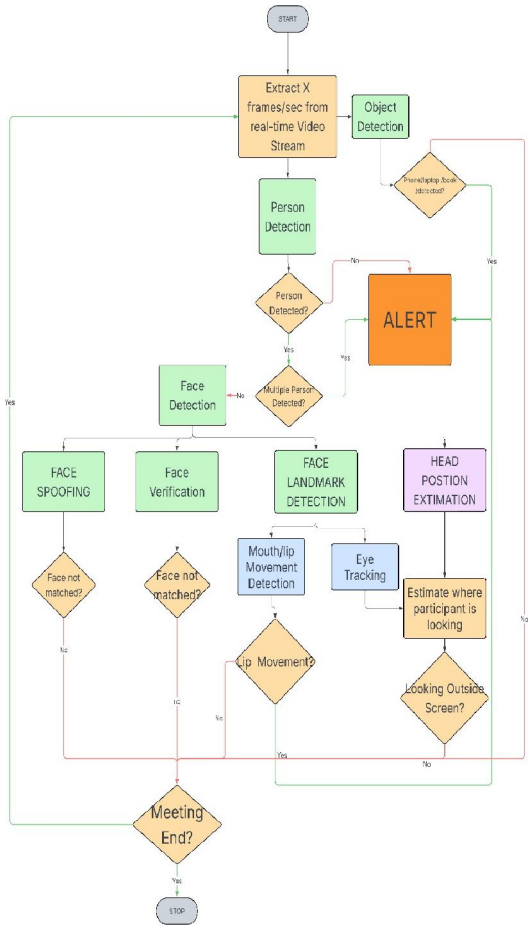


Fig. 3. Architecture of the proposed Online Proctoring System

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

Person Detection and Counting

We used OpenCV's [11] YOLOv8 object detector for detect- ing and counting the number of people in the frame. If no one or more than one person is detected for more than 10 consecu- tive frames, then the examinee is said to be cheating. Outputs from the Person Detection and Counting module are visible in Fig. 4. and Fig. 5.

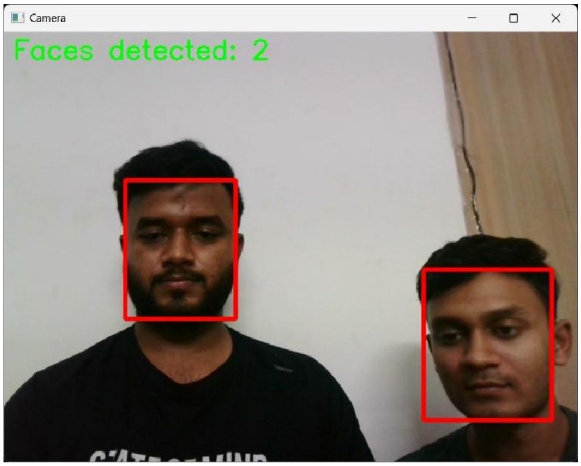


Fig. 4. Output from Person Detection and Counting module for frame with double person.

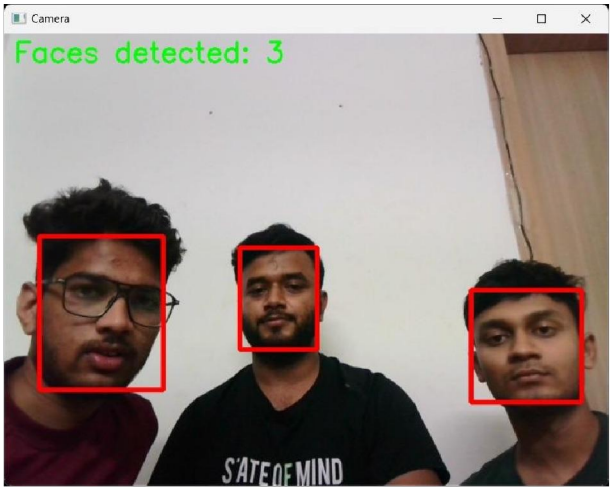


Fig. 5. Output from Person Detection and Counting module for frame with multiple person.

Object Detection

OpenCV's YOLOv8 object detector was also used for finding any instances of banned items including mobile phones, lap- tops, TV, and books. If one or more than one instance of any banned item is detected for more than 10 consecutive frames, then the examinee is said to be cheating. Outputs from the Object Detection module are shown in Fig. 6. and Fig. 7

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

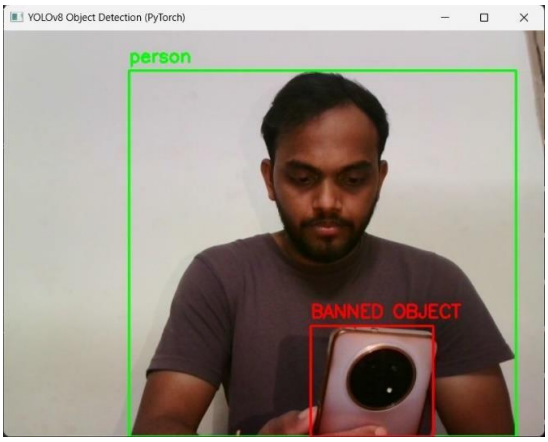
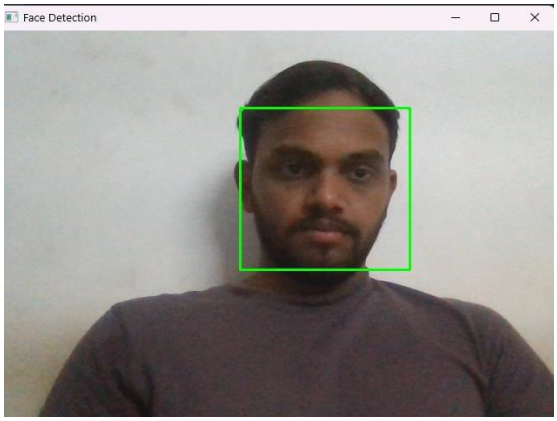


Fig. 7. Output from Object Detection module for frame with banned object.

Face Detection

We used OpenCV's DNN (Deep Neural Network) module to find the examinee's face in the frame. The face detector is based on the Single Shot Detector (SSD) framework with a ResNet base network. Output from the Face Detection module is shown in Fig. 8.



Authentication

Face Verification

We used Dlib's [12] face verification model to get the exam- inee's name. The model uses a pre-trained Resnet50 CNN model to extract a 128D feature vector from all facial images in the database. Then the model uses the same steps on the de- tected examinee's face to extract a 128D feature vector. After that, Euclidean distance is calculated between the two feature vectors. If the distance is below a certain threshold, then both faces are assumed to be the same. Dlib's default threshold of 0.6 was used for face verification. Outputs from the Face Verification module are illustrated in Fig. 9. and Fig. 10.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

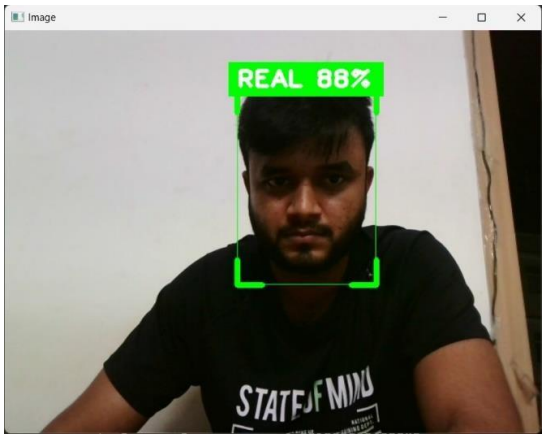


Fig. 9. Output from Face Verification module for frame with valid examinee.

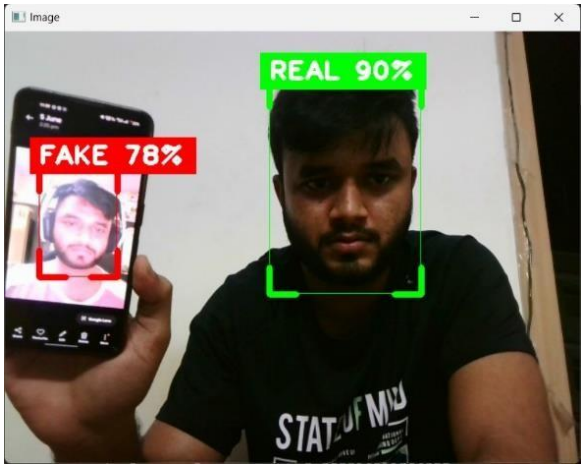


Fig. 10. Output from Face Verification module for frame with invalid examinee

To identify whether the examinee is real or a photograph, we implemented face spoofing functionality. After capturing the examinee's face image using the Face Detection module, it is further converted into YCrCb and CIE L*u*v* color spaces using OpenCV. Later, histograms are calculated from both the color spaces and concatenated together. The concate- nated histogram is sent to Scikit-learn's [3] ExtraTreesClas- sifier model for classifying face into real/spoof. If the face is classified as a spoof for more than 10 consecutive frames, then the examinee is said to be cheating. Outputs from the Face Spoofing module are shown in Fig. 12.

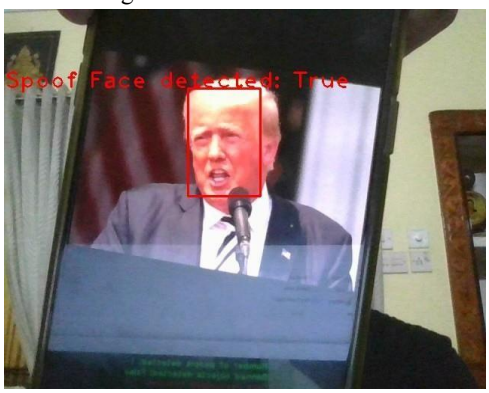


Fig. 12. Output from Face Spoofing module for frame with spoof face.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

Abnormal Behavior Detection

Head Pose Estimation

We trained a lightweight head pose estimation model that can achieve real-time performance in a system with low compu- tational power. Our method can predict accurate pitch, yaw, roll angles of a person directly from the face crop without the requirement of landmark or depth maps. We decided to train a pose estimation network from scratch instead of using land- mark because we believe that deep networks have large ad- vantages compared to landmark-to-pose methods due to the following reasons:

• Deep networks are not dependent on the head model chosen, the landmark detection method, the subset of points used for alignment of the head model, or the optimization method used for aligning 2D to 3D points [13].

• They always output a pose prediction which is not the case for the latter method when the landmark detection method fails especially for extreme poses [13].

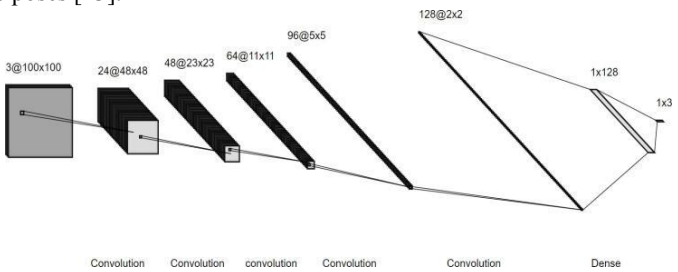


Fig. 13. Convolutional neural network architecture of pro- posed head pose estimation model.

Our lightweight regression model architecture is illus- trated in Fig. 13. We used a 5x5 convolution layer with stride 2 for the first layer followed by 4 3x3 convolution layers with stride 2 for feature extraction. The feature extraction is followed by a dense layer with 128 nodes and the output regression layer with 3 nodes. For all layers except the last convolution layer and regression network, we used the ReLU activation function. For the last convolution layer before Flat- ten layer and the following dense layers, we used the tanh activation function. Finally, for the output layer, we used the linear activation function. Fig. 14. illustrates the output from the Head Pose Estimation module.

Fig. 14. Output from Head Pose Estimation module. 3D (red, green and blue) vectors are used to illustrate the predicted head pose angles (pitch, yaw and roll)

Eye Tracking

We used Dlib's pre-trained network for detecting and predicting 68 facial landmarks on the examinee's face. Left eye is de- fined by the following landmarks - 36,37,38,39,40,41. Right eye is defined by the following landmarks - 42,43,44,45,46,47. First, we segmented the eye regions by using a mask. Then we applied binary thresholding on the eye regions to sepa- rate the eyeballs from the rest of the eye regions. Eyeballs become black and the rest regions stay

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

white. Then a vertical separator was created at the middle of each eye. Finally, to determine if the examinee is looking left or right, we defined an eye-tracking ratio as:

AvgETR = RightEyeETR + LeftEyeETR / 2

where,

RightEyeETR = No. of white pixels on left side/ No. of white pixels on right side

LeftEyeETR = No. of white pixels on left side / No. of white pixels on right side

After extensive trial and testing, we fixed the follow- ing thresholds for the AvgETR: ≤0.35 (looking outside the screen), 0.36 to 3.9 for the center (looking at the screen), ≥4 for left (looking outside screen). If the examinee is look- ing outside the screen for more than 10 consecutive frames, then the examinee is said to be cheating. Output from the Eye- tracking module is shown in Fig. 15.

Fig. 15. Output from Eye tracking module.

Mouth Movement Analysis

Lip region is defined by the following landmarks - 60, 61, 62, 63, 64, 65, 66, 67. To determine if the mouth is open or closed, we defined a lip aspect ratio as:

L.A.R = |P (62)–P (66)| / |P (60)–P (64)|

After extensive trial and testing, we fixed the threshold to be 0.1. If L.A.R>0.1, this means the mouth is open, else it is

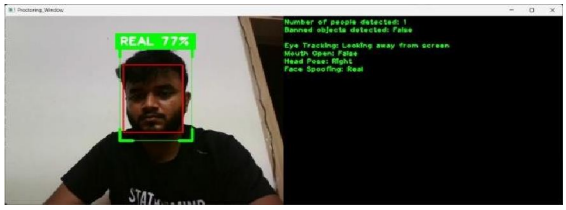


Fig. 16. Output from Mouth Movement Analysis module

Furthermore, to check if the person is speaking or not, we defined a buffer. If the examinee keeps his mouth open for more than 10 consecutive frames, then the activity is classi- fied as speaking and hence cheating. Outputs from the Mouth Movement Analysis module are shown in Fig. 16. and Fig. 17. for frame containing examinee with closed mouth.

IV. EXPERIMENTAL RESULTS

4.1. Head pose Training

Our model was trained on images from the Pandora dataset. The model takes a face crop rescaled to 100 x 100 as in- put. The face detection bounding box output is enlarged by 100% and the resulting bounding box is used to crop out the head region and pass that as input to the network. We used the Adam optimizer with a learning rate of 0.001 to train the model. The model was either trained for 100 epochs or was used with early stopping which monitored the validation loss with the patience of 20 epochs. The corresponding training loss and validation loss achieved for the best model is shown in Fig. 18. The model was trained using Mean Squared Error (MSE) as a loss function and evaluated on the test dataset us- ing Mean Average Error (MAE). To prevent the model from overfitting we used dropout regularization after each convo- lution and dense layer.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

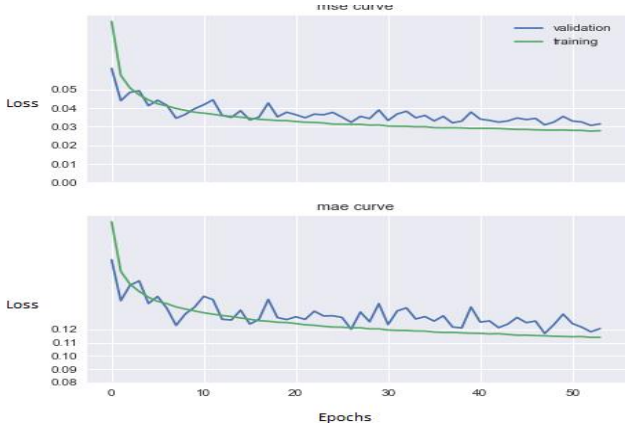


Fig. 18. Training and validation loss curves for Head Pose Estimation model.

4.2. Head pose Evaluation

We evaluated our lightweight head pose estimation model on the BIWI benchmark dataset. The dataset has around 15k RGB images with ground truth head pose angle values. The owners of the Pandora dataset have also provided the cropped face region for the BIWI dataset. So, we used this crop di- rectly without manually cropping the face region from the original benchmark dataset. Our lightweight head poses esti- mation model was able to achieve a better accuracy compared to famous landmark-based approaches and 3D dense model- based methods. Table 3 shows the comparison of results ob- tained by our model with Dlib [12] and 3DFFA [6] models. We took performance evaluation results of Dlib and 3DFFA from Hopenet's (state-of-the-art model) paper. Hopenet is a deep learning-based classification method for fine-grained head pose estimation. Even though the accuracy obtained by Hopenet is much higher than our lightweight model, we haven't used it in our online proctoring examination system. The Hopenet uses Resnet as the backbone feature extractor because of which, the computational complexity of the model is much higher and will not provide real-time performance for low-cost systems. Hence its performance is not compared with our lightweight model in Table 3. MAE is used to eval- uate the performance of all models.

Fig. 19. illustrates the predicted head pose vectors of our model on the BIWI benchmark dataset. This shows that our model was able to perform decently even for large head pose angles.

Table 3. Comparison of proposed head pose estimation model accuracy (MAE) with other available models

Model	Pitch	Yaw	Roll	MAE
Our model	11.000	13.927	7.471	10.799
Dlib	13.802	16.756	6.190	12.249
3DFFA	12.252	36.175	8.776	19.068

Fig. 19. Results by proposed head pose estimation model on BIWI benchmark dataset.

DOI: 10.48175/IJARSCT-27905

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

Feature	Tool Used	Why Used	Alternatives
People/Object Detection	YOLOv 3	Real- time, accurate multi-object	RCNN, SSD
Face Recognition	OpenCV + dlib	Accurate, easy integration	DeepFace
Eye Tracking	Dlib landmarks	Lightweight, precise EAR method	MediaPipe
Mouth Detection	Dlib + MAR	Fast, no extra model needed	Audio, CNNs
Spoof Detection	YCrCb/ LUV Histogra m	Quick liveness check via color	3D, Deep CNN

V. CONCLUSION AND FUTURE WORK

In this work, we proposed an Intelligent System that uses a webcam to monitor examinees during the online examination. The solution offers a comprehensive monitoring and analysis suite to prevent examinees from cheating in an online exam. Functionalities include user authentication and abnormal be- havior monitoring.

Currently, our pipeline runs at less than 30 frames per sec- ond (fps) because of the complexity of the models that we have used. In the future, we are planning to improve the fps by using lightweight models and proper optimization techniques like quantization. Moreover, the performance of the face spoofing model is not satisfactory. In the future, we are planning to replace it with a more accurate model. Currently, we are using an image processing based eye-tracking method. Later, we'll replace it with a deep learning based gaze estima- tion model that accurately estimates where the examinee is looking.in the video frame. (2) Object Detection - Find and report any instances of banned items in the video frame including mobile phones, laptops, TVs, and books. (3) Face Detection – Detect the face of the examinee in the video frame. (4) Face Verification - Match the ex- aminee's face in the video frame against a database of faces to authenticate the examinee. (5) Face Spoofing – Check whether the face of the examinee in the video frame is real or a photograph (spoof).

(6) Eye Track- ing - Track the eyeballs of the examinee during the exam and report if he/she is looking outside the screen.

(7) Mouth Movement Analysis - Find if the examinee speaks during the exam by checking if he/she opens the mouth for a certain duration.

REFERENCES

[1] Swathi Prathish, Kamal Bijlani, et al., "An intelligent system for online exam monitoring," in 2016 Interna- tional Conference on Information Science (ICIS). IEEE, 2016, pp. 138–143.

[2] Yousef Atoum, Liping Chen, Alex X Liu, Stephen DH Hsu, and Xiaoming Liu, "Automated online exam proctoring," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1609–1624, 2017.

[3] Senbo Hu, Xiao Jia, and Yingliang Fu, "Research on ab- normal behavior detection of online examination based on image information," in 2018 10th International Con- ference on Intelligent Human-Machine Systems and Cy- bernetics (IHMSC). IEEE, 2018, vol. 2, pp. 88–91.

[4] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.

[5] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li, "Face alignment in full pose range: A 3d total solution," IEEE Transactions on Pattern Analysis and Machine In- telligence, vol. 41, no. 1, pp. 78– 92, Jan 2019.

[6] Nataniel Ruiz, Eunji Chong, and James M. Rehg, "Fine- grained head pose estimation without keypoints," 2018.

[7] Yijun Zhou and James Gregson, "Whenet: Real- time fine-grained estimation for wide range head pose," 2020.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, June 2025
Impact Factor: 7.67

[8] Guido Borghi, Matteo Fabbri, Roberto Vezzani, Simone Calderara, and Rita Cucchiara, "Face- from-depth for head pose estimation on depth images," IEEE transac- tions on pattern analysis and machine intelligence, vol. 42, no. 3, pp. 596–609, 2018.
[9] Gabriele Fanelli, Matthias Dantone, Juergen Gall, An- drea Fossati, and Luc Van Gool, "Random forests for real time 3d face analysis," International journal of computer vision, vol. 101, no. 3, pp. 437– 458, 2013.
[10] Gary Bradski and Adrian Kaehler, "Opencv," Dr. Dobb's journal of software tools, vol. 3, 2000.
[11] S Sharma, Karthikeyan Shanmugasundaram, and Sathees Kumar Ramasamy, "Farec—cnn based efficient face recognition technique using dlib," in 2016 Interna- tional Conference on Advanced Communication Control and Computing Technologies (ICACCCT). IEEE, 2016, pp. 192–195.
[12] Nataniel Ruiz, Eunji Chong, and James M Rehg, "Fine- grained head pose estimation without keypoints," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2074– 2083